

Máster en Programación avanzada en Python para Big Data, Hacking y Machine Learning

Programación Python para BigData

LECCIÓN 06

VAEX y DASK

ÍNDICE

- ✓ Introducción
- ✓ Objetivos
- ✓ Dataset típico en aprendizaje de Big Data
- ✓ Alternativas para manejar grandes volúmenes de datos
- ✓ Conclusiones

INTRODUCCIÓN

En esta lección aprenderemos a trabajar con un par de herramientas útiles para aquellas situaciones que tenemos un gran número de filas en nuestro dataset.

OBJETIVOS

Al finalizar esta lección serás capaz de:

- 1 Conocer VAEX como herramienta para manejar grandes volúmenes de datos
- 2 Conocer DASK como herramienta para manejar grandes volúmenes de datos
- 3 Medir el tiempo de ejecución de una celda en Jupyter y la repercusión que tiene el ordenador que tienes en el mismo.

Dataset típico en aprendizaje de Big Data

The screenshot shows the official website of the NYC Taxi & Limousine Commission. The header includes the NYC logo, the text 'Comisión de taxis y limusinas', the phone number '311', and a search bar. Navigation links include 'Acerca de', 'Pasajeros', 'Conductores', 'Vehículos', 'Empresas', and 'TLC en línea'. A search bar is also present. Below the navigation bar, there are four main sections: 'Sobre TLC', 'Datos e investigación', 'Iniciativas TLC', and 'Comuníquese con TLC'. The 'Datos e investigación' section is highlighted, showing a sidebar with links to 'Datos', 'Programas piloto', 'Informes de la industria', and 'Libro de hechos'. The main content area is titled 'Datos de registro de viaje de TLC' and contains a paragraph describing the dataset.

NYC Comisión de taxis y limusinas 311 Buscar en todos los sitios web de NYC.gov

NYC
Taxi & Limousine Commission

한국어 ▶ Español | ▼ Tamaño del texto

🏠 **Acerca de** Pasajeros Conductores Vehículos Empresas TLC en línea

Search 🔍

Sobre TLC **Datos e investigación** **Iniciativas TLC** **Comuníquese con TLC**

Datos

Programas piloto

Informes de la industria

Libro de hechos

Datos de registro de viaje de TLC

Los registros de viaje en taxi amarillo y verde incluyen campos que capturan fechas / horas de recogida y devolución, lugares de recogida y devolución, distancias de viaje, tarifas detalladas, tipos de tarifas, tipos de pago y recuentos de pasajeros informados por el conductor. Los datos utilizados en los conjuntos de datos adjuntos fueron recopilados y proporcionados a la Comisión de Taxis y Limusinas de la Ciudad de Nueva York (TLC) por proveedores de tecnología autorizados en virtud de los Programas de mejora de pasajeros de taxis y librea (TPEP / LPEP). Los datos del viaje no fueron creados por TLC, y TLC no se responsabiliza de la exactitud de estos datos.

Alternativas para manejar grandes volúmenes de datos

- **Vaex**
Lo veremos en este tema.
<https://vaex.io/docs/datasets.html>
- **Dask**
Lo veremos en este tema
<https://dask.org/>
- **Rapids**
No lo veremos en este manual
<https://rapids.ai/>
- **Modin**
No lo veremos en este manual
<https://modin.readthedocs.io/en/latest/>
- **Ray**
No lo veremos en este manual
<https://ray.io/>
- **Koalas**
No lo veremos en este manual
<https://koalas.readthedocs.io/en/latest/>

CONCLUSIONES

1

VAEX es una gran alternativa cuando queremos hacer Big Data y su principal formato de datos es HDF5

2

DASK es una alternativa a VAEX que en esencia lo que hace es separar el dataset completo en muchos más pequeños pandas dataframes

3

Pandas es una gran herramienta pero quizá no la mejor cuando trabajamos con grandes volúmenes de datos



MUCHAS GRACIAS POR SU ATENCIÓN



jmpena@grupomainjobs.com



José Manuel Peña

<https://www.linkedin.com/in/jos%C3%A9-manuel-pe%C3%B1a-castro-7566b349/>



twitter.com/eiposgrados



facebook.com/eiposgrados



instagram.com/eiposgrados